Good Value:
Taking an evaluative approach in a field dominated by statistics and economics

a paper written for

2009 Australasian Evaluation Society Conference
Evidence *and* Evaluation

Canberra, Australia
31 August – 4 September 2009

Kathleen Palmer

Ministry of Research, Science and Technology
PO Box 5336 Wellington
New Zealand

email kathleen.palmer@morst.govt.nz
+64 4 917 3053

Keywords
Utilisation-focussed evaluation, Realist evaluation, Research and development,

The views expressed in this paper are those of the author only and do not necessarily represent the views of the Ministry of Research, Science and Technology. Any errors are due entirely to the author.

ABSTRACT

This presentation reflects on a utilisation-focussed evaluation that applied a realist evaluation approach in a traditionally economics and statistics domain.

The evaluation the New Zealand R&D tax credit benefitted from close collaboration between the science, revenue, finance, statistics and commerce agencies. Members from these agencies provided expertise, peer review and acted as 'critical evaluators'.

In 2008, baseline surveys were commissioned to gauge the quality of firm management of their research and development activities. These studies were to be followed up by monitoring surveys at two-year intervals. Each of the three surveys had different but related questions and samples, and importantly, were carried out by people with diverse strengths and interests.

An international business consultancy firm ran the early survey and in-depth case studies; an academic research unit conducted the second survey (of small and medium enterprises); and the third was a panel of questions in the national biennial survey of research and development in New Zealand.

The context for the evaluation underwent an extensive change late in 2008, and the evaluation programme stopped. Rather than leave the baseline survey work to languish, additional value has been extracted. For example, policy thinking around business R&D, our understanding of our national R&D survey and our R&D advice to firms has benefitted. This outcome is a consequence of taking a deliberate utilisation approach from the conception of the programme.

An implication for evaluation programme management is that the cross-discipline workspace and a proactive approach provide fertile ground for applying different tools and approaches to evaluation.

Author Kathleen Palmer

Email kathleen.palmer@morst.govt.nz

1 INTRODUCTION

The New Zealand government announced in Budget 2007 that it would introduce a tax credit for research and development (R&D tax credit) from the start of the 2008/09 income year. In late 2008, the government repealed the R&D tax credit legislation, and so the R&D tax credit ran for one full year.

The R&D tax credit was a new economic policy instrument. The credit was intended to encourage New Zealand businesses to invest more in R&D, to innovate and develop improved products and processes.[1] Businesses would submit a self-assessed claim for their R&D expenses and get a credit of 15 percent on any eligible expenditure.

There were multiple stakeholders, and some expectation that the evaluation would measure the economic concepts of deadweight and additionality. This evaluation needed to be robust and pragmatic.

The Ministry of Research, Science and Technology, drawing on the expertise of an interagency steering group, designed a five-year evaluation programme. Members of the steering group represented the policy, operations and evaluation sections of Inland Revenue, as well as the Treasury, the Ministry of Economic Development, and Statistics New Zealand.

The evaluation was internationally significant as it was designed alongside (rather than after) the tax credit, and drew on current international knowledge of R&D incentive evaluations.[2] The evaluative framework addressed each step in a simple logic model (intervention logic) of the R&D tax credit, and we expected that impacts might be observed by 2011 – the evaluation programme's planned finish date.

In early 2009, the evaluation was closed. The work commissioned for the evaluation of the R&D tax credit now contributes to the wider context of business R&D.

This presentation first describes the role of the steering group, looking particularly at the utilisation focus of the evaluation and the critical evaluator role. The presentation then looks at the overall design of the baseline surveys, and their realist evaluation features. The results of the baseline work is summarised. Finally, use of the survey findings in the wider business R&D context is illustrated.

[1] Research & Development tax credit questions and answers http://www.morst.govt.nz/publications/a-z/r/rd-qa/ and Research & Development tax credit fact sheet http://www.morst.govt.nz/publications/a-z/r/rd-fact-sheet/

[2] See CREST (2006) http://ec.europa.eu/invest-in-research/pdf/download_en/final_report_060306.pdf

STEERING GROUP AND UTILISATION FOCUS

The utilisation-focused approach[3] is explicitly geared to ensure that programme evaluations make an impact. The approach is pragmatic and ubiquitous, and a range of methodologies can be used. Elegance of design and technical excellence are of less importance than achieving meaningful use of the findings. In a utilisation-focused evaluation, the evaluator identifies the multiple and varied perspectives and interests that should be represented in the study.

In the evaluation of the R&D tax credit, the members of the steering group represented these different perspectives and interests. The different perspectives and interests of the steering group included: running a world class evaluation programme; maintaining a focus on R&D activity and its correct classification; assessing risks to revenue; identifying risks to data quality; technical correctness in survey design and statistical methods, and developing economic theory.

These perspectives provided a creative tension during the evaluation's design period. The economists noted that a set of perception surveys using a Likert scale would not be satisfactory; an experimental counterfactual design with randomised control and treatment groups was not practical as the intervention was a national one; and an econometric design using 'difference in difference' would fail the simplicity requirement and probably have significant methodological issues. The data people noted that national R&D statics would be weakened by methodology changes over the years, and our data and methods need to be able to refute critique – informed or otherwise. These discussions with stakeholders informed the choice to use a reflexive design[4] informed by realist evaluation.

The steering group members also had a valued role as 'critical evaluator'.[5] At one point we met to tackle the Frascati definition and tax legislation line by line because one member thought that differences in the wording of the two definitions might act as a block to our evaluation's integrity. We found that the differences, other than the obvious exclusions, were insignificant at the level of analysis we required for the evaluation. Our objective review of the definitions helped guard against 'group think'[6]; and the results of the review have been useful in reassuring other stakeholders.

The cognitive diversity of the group was maintained throughout the project and the members provided a valuable sounding board and 'fresh eyes' on survey designs and preliminary interpretation of results. Group members also peer reviewed the survey reports and programme documents.

---

[3] Stufflebeam, D.L. and A.J. Shinkfiend. (2007) Evaluation Theory, Models, & Applications. San Francisco: John Wiley & Sons.

[4] Reflexive design is a quasi-experimental approach where the counterfactual is constructed on the basis of the situation before and after the intervention. This design is particularly useful in evaluations of nationwide policies were there is no scope for a control group.

[5] A critical evaluator is a role taken on by members of a group as a precaution against development of the harmful aspects of 'group think'.

[6] Group think is a term coined by social psychologist Irving Janis (1972) and occurs when a group makes faulty decisions because group pressures lead to a deterioration of "mental efficiency, reality testing, and moral judgment".

BASELINE SURVEY DESIGN AND REALIST EVALUATION FEATURES

The purpose of a realistic evaluation is to establish whether there is an unequivocal causal relationship between a programme and its outcome. It identifies weather the mechanisms triggered by an intervention were intended, and observes precisely what happens in the space between input and outcome. Realistic evaluation is interested in how the context affects outcomes of a policy intervention. [7]

We needed to know how the R&D tax credit would affect R&D expenditure, and how it would affect the ability of firms to report their R&D.

We needed to know a great deal more about the under lying mechanisms firms used to produce R&D expenditure data. This was because firms had to measure their R&D in order to put in a claim, and because our output measure of business expenditure on R&D reported in national statistics [8&9] relied on firm assessment of their R&D expenditure.

There are three mechanisms for increasing reported R&D. Firms may:

- increase the amount of R&D they fund or perform – i.e. act as the policy intended
- improve the management of their R&D to complete claim forms accurately
- reclassify non-R&D work that could fit within the rules to increase their claim.

No prior significant work had been done in New Zealand on the accounting practices of firms for reporting R&D. We developed the following set of questions that formed the framework for each of the three surveys:

- How good is the definition of R&D used by firms when reporting R&D?
- How well do firms document their R&D activities?
- How well developed are firms' systems for recording R&D expenditure?
- What difficulties do firms have in externally reporting of R&D?
- What expectations do firms have for their R&D activity in the next financial year?

A set of 3-5 objective 'descriptors' followed each question (see Table 1). The firms chose the descriptor that best described what the firm was doing.

We also needed to know whether R&D activity and R&D management differed by contextual factors (another feature of realist evaluation) such as firm size, industry and location, and so each survey also collected data on a range of firm attributes.

---

[7] Pawson, R. and Tilley, N. (1997) *Realistic Evaluation.* London: Sage

[8] Ministry of Research, Science and Technology (1991) New Zealand Research and Development Statistics Business Enterprise Sector, 1989/90.

[9] www.morst.govt.nz/publications/statistics or http://www.stats.govt.nz/economy/innovation-and-science/researchanddevelopment.htm

The Ministry of Research, Science and Technology commissioned three surveys that ran during the first half of the year that of the credit before firms could complete a claim.

The first study was the smallest, with the closest relationship between researcher and participant. It provided early data (August 2008) on 121 firms and in depth case studies of 13 firms. The case studies verified that self-reporting was sufficiently accurate for our purposes. The multi-national company that ran this survey used more business-like language and business process questions, and referred to the R&D tax credit.

The second study, was larger and focussed on small and medium sized enterprises. It brought in data from 252 R&D-active firms. The university research centre that ran this study specialises in small and medium enterprise research. The survey used more personal and general language, and included questions about innovation. This survey ran during the election period and so did not draw attention to the R&D tax credit.

The third and largest study identified 600+ R&D active firms, was technically better, but asked fewer questions that the others. It was in the form of a panel of questions in the New Zealand R&D survey which follows the OECD methodology guidelines.[10] This survey underwent formal cognitive testing and used the most neutral and specific language. It had the highest response rate (so much lower non-response bias), and had carefully calculated weighting data for each firm from which we could draw a national picture.

An example of how the questions were presented across the three surveys is shown in Table 1. The different backgrounds and skills of the researchers enriched the studies.

TABLE 1 EXAMPLE QUESTION

STUDY	QUESTION	DESCRIPTORS
Medium-firm study	Which of the following best describes the systems your firm has to record R&D expenditure?	• No formal recording of R&D expenditure • R&D expenditure recorded, but not separately accounted for in financial statements • R&D expenditure recorded and separately accounted for in financial statements
Small-firm study	Which of the following best describes how you record your R&D expenditure?	• I keep separate records of R&D expenditure at each project stage • I keep separate records of R&D expenditure for the overall project, but not at each project stage • I keep no separate records of R&D expenditure, but we have timesheets that can be used to figure out time spent • I keep no records at all that could be used to identify R&D expenses
National R&D survey	Which of the following best describes the systems this organisation currently has for reporting expenditure on in-house or external R&D?	• Separate reporting of expenditure and costing methods at each R&D project stage • Separate reporting of R&D and non-R&D related expenditure • R&D and non-R&D related expenditure recorded together

[10] Benoit, G. (2008) The making of statistical standards: The OECD and the Frascati Manual, 1962-2002. Project on the History and Sociology of STI Statistics Working Paper No. 39 http://www.csiic.ca/

		under common expense categories

SUMMARY OF FINDNIGS

The three surveys were designed for different, but complementary purposes. They did not have the technical elegance that would allow data to be pooled or be compared directly. The findings showed strong congruence with remarkable similarity in places.

Definition of R&D

Generally, familiarity with a formal definition of R&D was low. Most small firms had an informal understanding of R&D. Fifty-seven percent of R&D-active small firms reported that they go with their 'gut feeling' when deciding whether an activity was R&D or not.

Reporting of expenditure in the national R&D survey was not always based on the Frascati definition of R&D. Of the firms in the medium-firm study that had reported expenditure to the R&D survey previously, 17 percent had applied the R&D definition provided on the R&D survey. A larger proportion (43 %) said that they adapted their figures to the supplied definition if possible; otherwise, they applied their own definition. The financial reporting standard (FRS-13) was used by 23 percent, and the other 17 reported that they used their own (non-Frascati or FRS-13) definition.

Record keeping

Comprehensive **project documentation** was reported by about 15 percent of firms in the medium- and small-firm studies. For each of the three studies, about half of the firms had some formal, or more comprehensive documents for their R&D projects, and the other half had no specific process for documenting R&D activity.

For the medium and small firm samples, the proportion of firms that were clearly meeting Frascati standards of a project budget and **expenditure records** was 10 percent. The national R&D survey, with its lower non-response bias suggests that the national percentage is lower than 10 percent.

Barriers

Understanding the definition of R&D was the most significant issue faced when reporting R&D activities and expenditure by participants in the medium-firm sample (53%) and the national R&D survey (29%).

A much higher number of firms reported no difficulties in the later national R&D survey (38%) than in the earlier medium-firm study (6%).

Next financial year

Forty-four percent of the medium-firm sample, 65 percent of the small-firm sample and 33 percent of the national R&D survey expected no change in the amount of R&D they would do or fund in the next financial year.

In addition to the descriptive statistics prepared for each evaluative question, we used non-parametric statistics to address some long held assumptions about how well firms understood the Frascati definition of R&D (Table 2 lines 1-3) and about their record keeping (Table 2 lines 4 & 5). The commonly known statistical analysis[11] provided some certainty to the results, which supported some, but not all assumptions. This work illustrates the value of investigating assumptions about context were feasible.

TABLE 2 NULL HYPOTHESES AND SUMMARY RESULTS

	NULL HYPOTHESIS	CHI-SQUARE RESULT
1	Large and SMEs firms do not differ in their understanding of R&D definition	Accepted: No size effect for difficulty with the definition of R&D, SMEs and large firms are similar.
2	Service firms and manufacturing firms do not differ in their understanding of R&D definition	Accepted: No industry effect for difficulty with the definition of R&D – services and manufacturing are similar. In the medium-firm study, there was no overall pattern of industry sector dominance in firm responses. In the small-firm study, "I go with my gut feeling" was more common in the service sector than for manufacturers but there was no difference in the level of personal study to learn about R&D between sectors.
3	Small science research firms do not differ from small non-science research firms in their understanding of R&D definition	Accepted: No effect for difficulty with the R&D definition. But small science research firms had better documentation and expenditure records. Proxy data from the small-firm study suggest that firms where R&D is 'core to business' are better managers of R&D documents and expenditure.
4	SMEs and large firms do not differ in the quality of their project documents	Rejected: Large firms have better documents. In the small-firm study, small and micro firms did not differ in their quality of R&D documentation.
5	SMEs and large firms do not differ in the quality of their expenditure records	Rejected: Large firms have better records. In the small-firm study, small and micro firms did not differ in their quality R&D expenditure records

[11] Nonparametric statistical procedures are designed for ordinal or nominal data. The statistical test used to test the hypotheses was the non-parametric Pearson's Chi-square test of independence. The Person's Chi-square test for independence tests whether observations from two or more actual samples (in this study size classes or industry types) differ from each other. The statistical testing used data in the form of counts. The number of degrees of freedom was one and the p value (probability value) was 0.05. This means that there was a five percent chance that a null hypothesis is rejected when it is in fact true.

USE OF THE SURVEY FINDINGS IN A WIDER CONTEXT

The baseline studies were initially designed as the reference point, against which we would measure how much change has occurred as a result of the intervention.

This section provides examples of how we got value from the reference point itself.

Evaluation design updated: The first survey to be completed was accompanied by a set of 13 case studies. These case studies demonstrated that the self-report survey results were adequate for our needs. As a result, we did not fund additional case studies.

Business R&D information updated: The study of small and medium enterprises found that firms get good ideas or information for their innovation activities (includes R&D) from customers, other businesses and employees. Firm owners also go to books, journals and the internet. From their support structure, innovating firms considered industry associations to be the most important sources of information. The results showed that accountants, lawyers and bankers, who generally play an important role as a source of information, were less important when it came to innovation. Good communication between our evaluation and business-facing people meant that the glossy publication on making R&D work for business[12] was updated with the new information in time for publishing.

R&D policy thinking stimulated: The evaluation reports hold material that could be used by policy analysts if the information was extracted and further refined. To help with the transfer of ideas from the evaluation to the policy setting, I prepared a specialist report to synchronise with the policy work plan. This report has had some success. For example, the surveys found that firms were struggling to apply a formal definition of R&D, and that there was a lot fine-tuning involved in fully understanding the Frascati or tax credit definition of R&D. This finding has contributed to a policy question: How can government incentivise business R&D effectively and fairly, but at a lower transaction cost that the R&D tax credit was likely to extract?

National R&D statistics study ongoing: Many firms reported that they were doing R&D, but were not keeping project plans or not setting a budget or keeping expenditure records. These patterns may affect the integrity of international comparisons of R&D statistics. We do not know whether New Zealand is better or worse than other countries. However, like New Zealand other countries are beginning to raise questions about national patterns of reporting R&D.

Evaluation capacity and project management respected: The evaluation involved stakeholders from the outset. The steering group's specifications were married with evaluation theory to design a pragmatic and robust programme. Results were shared early, researcher opinions' were valued and experts' contributions applied. As a result, we produced a credible piece of work that maintained its value, and have lifted our capacity in both evaluation design and the management of large evaluation projects.

[12] Edmond, K; (2009) Idea to Impact: Making R&D work for your business. Wellington: Ministry of Research, Science and Technology.

8 CONCLUSIONS

The cross-discipline and proactive features of the evaluation illustrate Patton's 'utilisation-focussed evaluation'. Pawson and Tilley's 'realistic evaluation' approach resulted in crucial insight into the potential for mechanisms related to R&D i.e. its management to have a substantial impact on outcomes.

The steering group has provided input that helped in the design and management of the evaluation. The different strengths and interests, cooperation, and cognitive separation of the group members provided benefits to the evaluation programme.

The use of different groups to run the surveys also added flavour and interest. The designs were pragmatic and attentive to the nature of their participants.

Even in a changed environment, there were real benefits from the evaluation. This value was a consequence of the proactive approach to setting up and managing the evaluation, cross-discipline input from stakeholders, adept selection of tools and approaches, and the resulting utility of the findings to different audiences.

To obtain copies of source reports go to

http://www.morst.govt.nz/publications/evaluations/rd-tax-credit/